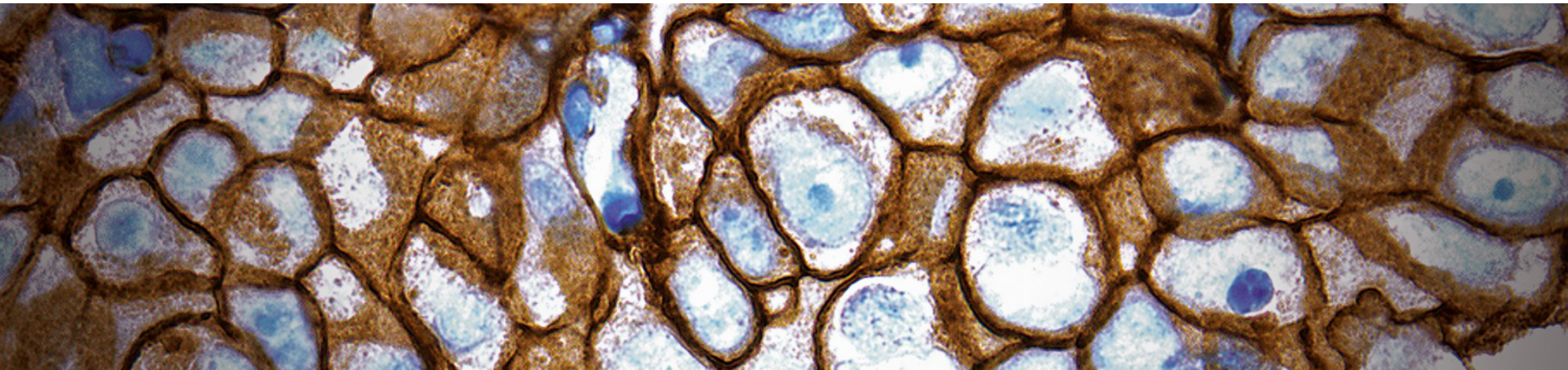
My “Big Data” experience

– getting my feet wet in the puddle and still treading water 2 years later

28th June 2013

Rob Walls

Healthcare Data Analytics



Abstract

"Big Data" - Electronic Health Records (EHR) and Health Insurance Claims data are growing exponentially every day. Regulatory focus has widened to ensure approved pharmaceutical products are efficacious and safe in broader populations in the Real World. An area of extreme interest in today's ever changing and increasingly competitive pharmaceutical industry but what exactly is it? Where does it come from? What does it contain? How can we use this type of data to inform on the safety of marketed products and during the development of pharmaceuticals?

My first two years of exposure since clinical trial programming have meant many perception shifts. I will explain data structures, touch on some SAS tips and tricks, and describe the differences between EHR and Claims data. Most importantly I will show how the right analysis can help effectively inform and affect decision making across all stages of the molecule's development lifecycle and marketing.

Introduction

- Data – Clinical vs. Real World
- Studies – Clinical vs. Real World
- Shifts in perception
- SAS Efficiency Considerations
- Electronic Health Records Data
- Insurance Claims Data
- Data Usage
  - Clinical Trial Patient Recruitment
  - Health Economics
  - Safety Input
  - Molecule Development
  - Risk Estimation
  - Comparative Effectiveness
- Conclusion

Data – Clinical Trials vs. Real World

Clinical Trials	Real World
Primary purpose data	Secondary purpose data
Clean data (relatively) - Variable mappings and formats are predefined - By reporting events the data should be of very high quality	Unclean data - Data filled with inconsistencies and irregularities
Consistent expected data structure (CDISC, company specific, etc.)	Structures are dictated by vendors dependent on primary purpose.
Relatively small data sets - Max ~500,000 observations	Huge, unwieldy data sets - Max ~3,500,000,000 observations

Studies – Clinical Trials vs. Real World

Clinical Trials	Real World
Hypothesis testing -studies are planned and laid out according to the strict guidelines in the Protocol to provide evidence behind the science	Hypothesis generation -studies are carried out in an investigatory manner which can later evolve dependent on the results being seen in the data
Full control over how each data point is generated	Each database has its own idiosyncrasies which can effect it's suitability for use. It can be unclear how some variables are derived so often clarification is required from the vendors
Clearly defined timelines around the delivery of data and results	Timelines are often variable, shifting based on need
Support of a molecule through distinct stages of development (Phase I-IV)	Provide support of molecules/ therapeutic areas across all aspects of the clinical development lifecycle

Shifts in perception

- ‘Clean’ data = Good data
  - → EHR/Claims data = Unclean data, but this is still a rich data source
  - → The larger the population the less skewing of the results due to a few missing or incomplete variable values
  - → Workarounds may be possible
- A clinical trial is a clinical trial. One TA, one phase, one molecule at a time
  - → Studies can come in for support of different therapeutic areas, molecules, or different stages of development.
- I know how to code efficiently…….
  - → There must be a real focus on efficiency and efficiency techniques.

SAS Efficiency Considerations

- Taking into account the I/O performance
- Effective/intelligent use of the Logical Program Data Vector (LPDV)
- Indexes
- Hash objects (if appropriate)
- Buffers and buffer sizes
- How best to pull data from base database
  - Parallel pulls of data (for example if stored by years)
  - Data set options on the SET statement
    - limit the data to ONLY read in what is required
    - when using an indexed variable the syntax in a where clause can often force SAS to ignore the index (without informing the user)
  - Variable length and type
- Merge efficiencies – Which technique works best for the data that you have?

Electronic Health Records Data

- Primary purpose is provision of healthcare to patients
- General practitioner databases
  - Clinical Practice Research Datalink [CPRD©] – (+50m)
- Healthcare Networks
  - General Electric - Electronic Medical Records [GE-EMR©] – (+25m)
- All records to this particular provider will be captured
  - UK has a universal healthcare system
  - US has a healthcare network setup so patients can come and go

Electronic Health Records Data (cont.)

Pros	Cons
Specificity: -Labs -Lifestyle factors (BMI, Smoking status)	No payment information
Contains prescriptions written or drugs given during office visit	Contains prescriptions written
Prescription information is extensive	Type of data provider is key to understanding what data to expect to be able to find
Potentially we can have all data from a patient from birth to death	
True denominator can be derived by only considering ‘active’ patients	

Insurance Claims Data

- Primary purpose is to record payments of service
  - Optum Insight InVision Data Mart© [US] – (+32m)
  - Truven MarketScan© [US] – (+140m)
- All types of service will be captured while patients are enrolled into an insurance scheme
  - Claims will be paid across any healthcare networks regardless

Insurance Claims Data(cont.)

Pros	Cons
All services captured	Labs are only available for a subset of patients
Start and stop dates of enrollment in healthcare insurance plan so a true denominator can be returned	(Generally) No lifestyle factors
Can establish who is NOT using benefits	Cost information, where it exists, is imputted
Contains all medical diagnoses, medications that have been filled (dispensed) and procedures undergone from visits to all types of service	
Contains payment and charge information.	

Data usage – Clinical Trial Patient Recruitment

- Generation of heat maps showing a geographical representation of disease distribution.
- Quantification of potential patient recruitment overlaps
  - Two trials for the same indication but slightly different subpopulations
  - Data can provide an overview of the scale of the overlap
  - Allows insights into whether studies will be in direct competition with one another
  - Can influence site selection for both studies

Data Usage – Safety Input

- Safety Surveillance
  - Proactive monitoring of marketed compounds by drug companies
- Clinical trial submissions
  - Provide context into clinical trial data being supplied to regulators
  - Investigations can be carried out to provide context into potentially concerning safety aspects of clinical trials
- Investigations of off-label treatment patterns
  - New requirement from the EMA which will feed into the Periodic Benefit Risk Estimation Report (PBRER)
  - How to define ‘off-label’? Not always as straight forward as it may sound.
  - Some prescriptions may be attributable to the on-label diseases, however, this can be ‘fudged’ by medical practitioners for reimbursement purposes. How do we identify this?
  - Once we’ve identified the off-label usage, how do we work backwards to identify what it is being used off-label to treat?

Data Usage – Health Economics

- Analyses can be done (using the insurance claims data sources) into the charges and expenditures relating to certain treatments
- Funding bodies (NICE, etc.) may not be willing to pay for your drug
  - Cost/benefit ratios can be analyzed and provided
  - Comparative studies can be performed into competitor compounds
  - Evidence can be gathered to help build an argument

Data Usage – Molecule Development

- Clinical trials are long and costly, therefore it is important when making decisions on the molecule level to have a good context on the cost of trials versus the potential risk/ benefit to be had for patients
  - Inform on decisions to end a molecules development early thereby saving money
  - Enable the continuance of development perhaps where a unique patient value is seen
  - Clearly pre-defining the disease area under investigation. The more you know about the intended recipients of your drug the more informed and directed the clinical trials will become.

Data Usage – Risk Estimations

- The evaluation of ‘Risk’ is vital in the drug development lifecycle and feeds into everything that we now do
  - Risk/benefit profiles must be build for each molecule and these now feed into the Risk Management Plan (RMP)
  - Monitoring of risks associated within populations taking certain compounds or with certain diseases in order to gauge potential impact.
  - Given what is known about our molecule, is it ethical to treat patients with a certain history profile?
  - Risk plays a major part in all safety analyses

Drug Usage – Comparative Effectiveness

- On the market research comparing competing drugs for
  - Efficacy
  - Costs
  - Risk/benefit ratios
  - Regional variations in usage
- Allows companies to better target and place their drugs on the market to better benefit patients
- Can provide evidence to payers in order to back up arguments for inclusion of a drug into the approved usage lists (formularies).

Conclusion

- The world of EHR and Claims data is vastly different to the clinical trials
- Dirty data, not necessarily fit for purpose data
  - Large, unwieldy and complex, but with some manipulation can be made to be as ‘fit for purpose’ as possible.
  - What you see is not always what you get
- Gives the ability to answer and provide insights into many important questions across every aspect of the clinical life cycle
- Data is expanding. Each year the amount of data we are seeing increases in both complexity and volume.
  - Vendors are bringing in more practices/insurers and are linking their data to provide richer, more robust databases. The aim is to make it more suitable for research use while retaining it’s original ‘primary’ purpose. Links such as state registries (births/deaths, cancer registries, hospital encounters, labs) are already being seen.
  - Great for the scope of analyses that can be performed, but a big ask for programmers across the field.

Conclusion

- This data is a very powerful tool and when used correctly can and is changing the way that clinical trials are managed
- It may even one day be possible to conduct clinical trials directly using data from these databases.
  - Studies are already being conducted using pre-existing clinical data to establish non-treatment arms in clinical trials, why not claims or EHR data?
  - See the EU's Innovative Medicines Initiative Electronic Health Records for Clinical Research – IMI EHR4CR

  https://www.imi.europa.eu/web/portal/home

  http://www.ehr4cr.eu/
- Even in the two years that I have been working in the area the scope of what we can and are achieving is constantly expanding. This is an evolving and exciting area to be working in at the moment so I'm going to keep trying to swim.


ANY QUESTIONS ?

Doing now what patients need next